

W53



EP0801344

Biblio

Desc

Claims

Page 1

Drawing

**esp@cenet**

## An apparatus for reallocating logical to physical disk devices using a storage controller and method of the same

Patent Number: ☐ EP0801344, A3

Publication date: 1997-10-15

Inventor(s): YAMAMOTO YASUTOMO (JP); SATOH TAKAO (JP); YAMAMOTO AKIRA (JP)

Applicant(s):: HITACHI LTD (JP)

Requested Patent: ☐ JP9274544

Application Number: EP19970105448 19970402

Priority Number(s): JP19960085370 19960408

IPC Classification: G06F3/06

EC Classification: G06F3/06M, G06F11/10M, G06F11/20L

Equivalents: ☐ US5956750

### Abstract

A storage controller (104) includes that it calculates an access frequency (500) of each logical disk (200); that it selects first logical disk device of which the access frequency exceeds a first predetermined value, the first logical disk device being allocated to a first physical disk device; that it selects a second logical disk device which has the access frequency equal to or less than a second predetermined value, the second logical disk device being allocated to a second physical disk device; and that it reallocates the first and second logical devices to the second

and the first physical disk devices, respectively.



Data supplied from the esp@cenet database - 12

(51)Int.Cl. <sup>*</sup>	国内整理番号	F I	技術表示箇所
G 0 6 F 3/06	5 4 0	G 0 6 F 3/06	5 4 0
	3 0 1		3 0 1 X
	3 0 2		3 0 2 E
	3 0 4		3 0 4 N
12/08	3 2 0		3 2 0
	7623-5B	12/08	請求項の数字 9 O L (全 14 頁)
		審査請求	未請求

21) 出願番号	特願平6-85370	(71) 出願人	000005108 株式会社日立製作所 東京都千代田区神田麩町四丁目6番地 山本 康友 株式会社日立製作所システム開発研究所内 神奈川県川崎市緑生区王禅寺1099番地 株 式会社日立製作所システム開発研究所内 山本 彰 神奈川県川崎市緑生区王禅寺1099番地 株 式会社日立製作所システム開発研究所内 佐藤 孝夫 神奈川県小田原市国府津2380番地 株式会 社日立製作所ストレージシステム事業部内 伊理士 有近 幹太郎 (74) 代理人
22) 出願日	平成8年(1996)4月8日	(72) 発明者	山本 康友 山本 彰 佐藤 孝夫

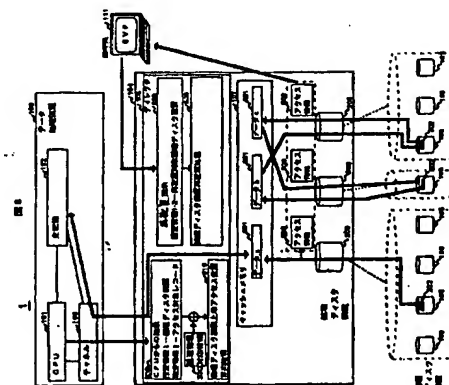
## 54) 【発明の名称】 配位側面装置

(57)【要約】

【課題】 アクセス性能を向上する。

【解決手段】 各論理ディスク装置200に対するアクセス情報500を採扱し、そのアクセス情報500をSVIP111を通じて保守員に提示する。保守員の所配置指示620があること、指示された2つの論理ディスク装置200の間で物理ディスク装置105を配置替えし、全200を連続的に格納し直す。

【効果】 アクセス頻度の高い論理ディスク装置をより高速な物理ディスク装置へ所配置することが出来る。シーケンシャルアクセスの比率的に高い論理ディスク装置をよりシーケンシャルアクセス性の高い物理ディスク装置へ所配置することが出来る。



## 【特許請求の範囲】

【請求項1】 データ処理装置が直接アクセスを行う論

【請求項2】 データ処理装置が直接アクセスを行う論理的記憶装置と実際にデータを記憶する物理的記憶装置とを対応付け、前記データ処理装置と前記物理的記憶装置との間のデータ転送を制御する記憶制御装置において、

物理的記憶装置へのアクセス情報を指標として採取するアクセス情報採取手段と、前記指標に基づいて前記論理的記憶装置を前記物理的記憶装置に再配置すると共に再配置の物理的記憶装置にデータを連続的に格納する論理的記憶装置再配置手段とを有することを特徴とする記憶制御装置。

【請求項3】 請求項2に記載の記憶制御装置において、前記アクセス情報が、前記データ処理装置から前記記憶装置へアクセス頻度情報へのアクセス頻度情報を含むことを特徴とする記憶制御装置。

請求項4 請求項2または請求項3に記載の記憶制御装置において、前記アクセス情報が、前記データ処理装置から前記論理的記憶装置へのアクセスパス情報を含むことを特徴とする記憶制御装置。

請求項5) 請求項1に記載の記憶制御装置において、前記指標が、前記論理的記憶装置に求められる信頼であることを特徴とする記憶制御装置。

請求項6) 請求項1から請求項5のいずれかに記載の記憶制御装置において、前記指標を保守員に提示する際、指示手段と、保守員からの再配置指示を受け付ける再配置指示手段とを具備したことを特徴とする記憶制御装置。

請求項7 請求項1から請求項5のいずれかに記載の記憶制御装置において、データ処理装置からの再配置指示を受け付ける再配置指示受付手段を具備したことを特徴とする記憶制御装置。

【請求項 8】 請求項 1 から請求項 5 のいずれかに記載の記憶制御装置において、前記指標に基づいて再配置の可否を決定する再配置要否決定手段を具備したことを特徴とする記憶制御装置。

**請求項9** 請求項8のいずれかに記載の記憶制御装置において、再配置中の論理的記憶装置にデータ処理装置からのアクセスがあったとき、再配置中の論理的記憶装置の再配置完了領域と再配置未完了領域とを識別し、前記アクセス位置が前記再配置完了領域なら、再配置先の論理的記憶装置にアクセスさせ、前記アクセス位置が前記再配置未完了領域なら、再配置前の論理的記憶装置にアクセスさせる。

セス位置が前記再配置未完領域ならば当該論理的記憶装置にアクセスさせるアクセス位置切替手段をさらに具備したことを特徴とする記憶制御装置。

50 [0001]

【発明の属する技術分野】本発明は、記憶制御装置に関し、さらに詳しくは、シーケンシャルアクセスの場合やランダムアクセスで低い場合でもアクセス性を向上させることができる記憶制御装置およびデータの種類性を向上させることができる記憶制御装置に関する。

特に、本発明は、ディスクアレイ向きの高機能ディスク装置、その高機能装置とディスク制御装置とにより構成される記憶装置サブシステム、およびその記憶装置サブシステムとデータ処理装置とにより構成される情報処理システムに有用である。

**[0002]**

【従来の技術】 シカゴのイリノイ大学で開かれた「ACM SIGMOD」会議において発表された論文「D. Patterson, G. Gibson, and R. H. Katz: A Case for Redundant Arrays of Inexpensive Disks (RAID), ACM SIGMOD Conference, Chicago, IL, (June 1988), pp. 109-116」は、ディスクアレ

100031

ディスク装置の一部をディスクキャッシュの如く用いる技術が開示されている。具体的には、ディスク装置を一般的にデータ格納するテンポラリ領域と最終的にデータを書き込む領域とに分け、更新データはバッティを生成せずに一旦テンポラリ領域に二重書きし、非同期にバッティ生成し、最終領域に書き込む。

**[0004]** -

95-68 (茂木他: Hot Mirroring を用いたディスプレイのディスプレイ故障時の性能評価, 1995年12月, 電気情報通信学会会報 Vol. 95-No. 407, pp. 19-2

0」には、ア

るRAIDレベルを動的に変更する技術が開示されている。具体的には、ディスク装置をRAID1構成の部とRAID5構成の部分に分け、ライトアクセスのあったデータを優先的にRAID1構成の部分に格納する

にうにデータ

格クセス頻度の高いデータはRAID1構成の部分に格納し、アクセス頻度の低いものはRAID5構成の部分に格納するように出来る。この技術によれば、記憶容量異なる物理ディスク装置やRAIDレベルの異なる物理ディスク装置を記憶装置サブシステム内で混在させる

とが可能であ

アアクセス頻度やアクセスパターンなどの指標に基づいた、任意の物理ディスク装置に格納することが出来る。また、アクセス頻度の高いデータを、より高速な物理ディスク装置に格納するように、動的に格納位置を変更することも出来る。なお、RAIDのディスクアレイ

は、データ処理装置からの書き込みデータに対して、その複製をミラーと呼ばれる副ディスク装置に書き込み、データの信頼性を確保する。元データが元のデータの複製であるため、元データ作成のオーバーヘッドが小さく、アクセス性能が良い。但し、物理的記憶装置の使用効率は、50%と低い。一方、RAID5のディスタレイは、データ処理装置からの複製の書き込みデータに対して、パリティと呼ばれる元データを作成する。パリティ作成時に更新前データと更新前パリティのリードが必要となり、元データ作成のオーバーヘッドが大きく、アクセス性能は悪い。但し、複製のデータに対して1つのパリティを作成するため、記憶装置の使用効率はRAID1に比べ高い。

【0005】

【発明が解決しようとする課題】上記従来技術では、アクセスするデータ単位でデータの格納位置の変更を行うため、データ処理装置が直接アクセスを行う論理ディスク装置上では逆接線データが、実際にデータに記憶する物理ディスク装置上では非逆接線となってしまう。このため、一連のデータをリード/ライトするシーケンシャルアクセスの場合、実際には順次データをまとめてリード/ライトできなくなり、アクセス性能の低下を招く問題点がある。

【0006】一方、上記報告「DE95-68」の従来技術では、ライトの度に、アクセス頻度が低いと判断したデータをRAID1構成の部分からRAID5構成の部分に移し、空いたRAID1構成の部分にライトデータを書き込み、アクセスパターンがランダムアクセスでヒット率が低い場合には、RAID1構成の部分に移したデータの多くは再びRAID5構成の部分に戻されることになる。このため、ヒット率が低い場合、アクセス性能の向上は期待できず、逆にデータを移す処理のオーバーヘッドがアクセス性能の低下を引き起こす問題点がある。

【0007】また、上記の従来技術では、データの信頼性の向上については全く考慮されていない問題点がある。

【0008】そこで、本発明の第1の目的は、シーケンシャルアクセスの場合やランダムアクセスでヒット率が低い場合でも、アクセス性能を向上させることが出来る記憶装置を提供することにある。また、本発明の第2の目的は、データの信頼性を向上させることが出来る記憶装置を提供することにある。

【0009】

【課題を解決するための手段】第1の観点では、本発明は、データ処理装置が直接アクセスを行う論理的記憶装置を実際にデータに記憶する物理的記憶装置に配置し、前記データ処理装置と前記物理的記憶装置の間のデータ転送を制御する記憶制御装置において、予め定めた指標に基づいて前記論理的記憶装置を前記物理的記憶装置に

再配置すると共に再配置先の物理的記憶装置にデータを連続的に格納する論理的記憶装置再配置手段を有する点とを特徴とする記憶制御装置を提供する。上記第1の観点による記憶制御装置では、アクセスするデータ単位でデータの格納位置の変更を行うのではなく、論理的記憶装置を単位として物理的記憶装置への再配置を行い、且つ、再配置先の物理的記憶装置にデータを連続的に格納する。従って、シーケンシャルアクセスの場合でも、アクセス性能を向上させることが出来る。また、ライトの度にデータの格納位置の変更を行うのではなく、予め定めた指標に基づいて前記再配置を行うから、ランダムアクセスでヒット率が低い場合でも、アクセス性能を向上させることが出来る。

【0010】第2の観点では、本発明は、データ処理装置が直接アクセスを行う論理的記憶装置と実際にデータに記憶する物理的記憶装置とを対応付け、前記データ処理装置と前記物理的記憶装置の間のデータ転送を制御する記憶制御装置において、前記データ転送の制御の適用中にデータ処理装置の論理的記憶装置へのアクセス情報を指標として採取するアクセス情報採取手段と、前記指標に基づいて前記論理的記憶装置を前記物理的記憶装置に再配置すると共に再配置先の物理的記憶装置にデータを連続的に格納する論理的記憶装置再配置手段とを有する点とを特徴とする記憶制御装置を提供する。上記第2の観点による記憶制御装置では、アクセスするデータ単位でデータの格納位置の変更を行うのではなく、論理的記憶装置を単位として物理的記憶装置への再配置を行い、且つ、再配置先の物理的記憶装置にデータを連続的に格納する。従って、シーケンシャルアクセスの場合でも、アクセス性能を向上させることが出来る。また、ライトの度にデータの格納位置の変更を行うのではなく、アクセス情報を採取し、それを統計的に利用して前記再配置を行うから、ランダムアクセスでヒット率が低い場合でも、アクセス性能を向上させることが出来る。

【0011】第3の観点では、本発明は、上記構成の記憶制御装置において、前記アクセス情報が、前記データ処理装置から前記論理的記憶装置へのアクセス頻度情報を含む点とを特徴とする記憶制御装置を提供する。上記第3の観点による記憶制御装置では、アクセス頻度の高い論理的記憶装置をより高速な物理的記憶装置へ再配置することが出来る。従って、アクセス性能を向上させることが出来る。

【0012】第4の観点では、本発明は、上記構成の記憶制御装置において、前記アクセス情報が、前記データ処理装置から前記論理的記憶装置へのアクセスパターン情報を含む点とを特徴とする記憶制御装置を提供する。上記第4の観点による記憶制御装置では、シーケンシャルアクセスの比率の高い論理的記憶装置をよりシーケンシャルアクセス性能の高い物理的記憶装置へ再配置することが出来る。従って、アクセス性能を向上させることが出来る。

出来る。

【0013】第5の観点では、本発明は、上記構成の記憶制御装置において、前記指標が、前記論理的記憶装置に求められる信頼性であることを特徴とする記憶制御装置を提供する。上記第5の観点による記憶制御装置では、信頼性が高いことが求められる論理的記憶装置をより信頼性の高い物理的記憶装置へ再配置することが出来る。従って、データの信頼性を向上させることが出来る。

【0014】第6の観点では、本発明は、上記構成の記憶制御装置において、前記指標を保守員に提示する指標提示手段と、保守員からの再配置指示を受け付ける再配置指示受付手段とを具備した点とを特徴とする記憶制御装置を提供する。上記第6の観点による記憶制御装置では、データ処理装置が再配置指示を入力できるように、保守員で判断可能な高度の条件下で前記再配置を行うことが出来る。

【0016】第8の観点では、本発明は、上記構成の記憶制御装置において、前記指標に基づいて再配置の要否を決定する再配置要否決定手段を具備した点とを特徴とする記憶制御装置を提供する。上記第8の観点による記憶制御装置では、記憶制御装置が再配置指示を自己決定するため、保守員やデータ処理装置に負担をかけることなく済む。

【0017】第9の観点では、本発明は、上記構成の記憶制御装置において、再配置中の論理的記憶装置にデータ処理装置からのアクセスがあったとき、再配置中の論理的記憶装置の再配置完了領域と再配置未完了領域とを識別し、前記アクセス位置が前記再配置完了領域ならば再配置先の論理的記憶装置にアクセスさせ、前記アクセス位置が前記再配置未完了領域ならば当該論理的記憶装置にアクセスさせるアクセス位置切替手段を具備したことを特徴とする記憶制御装置を提供する。上記第9の観点による記憶制御装置では、再配置中の論理的記憶装置の再配置完了領域と再配置未完了領域とを識別し、データ処理装置からのアクセス位置を切り替えるから、データ処理装置と物理的記憶装置の間のデータ転送を運用中に再配置を行うことが出来る。

【0018】

【発明の実施の形態】以下、本発明の実施形態を説明する。なお、これにより本発明が限定されるものではない。

【0019】-第1の実施形態-

第1の実施形態は、各論理ディスク装置のアクセス情報を記憶制御装置で採取し、SVP（サービスプロセッ

サ）を通じて保守員に提示し、このアクセス情報に基づいて保守員の再配置指示により、論理ディスク装置の物理ディスク装置への再配置を行うものである。

【0020】図1は、本発明の第1の実施形態にかかる記憶制御装置を含む情報処理システムのブロック図である。この情報処理システム1は、データ処理装置100と、記憶制御装置104と、1台以上の物理ディスク装置105と、SVP111とを接続して構成されている。

【0021】前記データ処理装置100は、CPU101と、主記憶102と、チャネル103とを有している。

【0022】前記記憶制御装置104は、1つ以上のデータレクタ106と、キャッシュメモリ107と、データレクタ108と、不揮発性メモリ109と、不揮発性メモリ管理情報400と、論理ディスク情報500とを有している。前記データレクタ106は、データ処理装置100のチャネル103と物理ディスク装置100のチャネル103とのデータ転送、データ処理装置100のチャネル103と前記キャッシュメモリ107との間のデータ転送および前記キャッシュメモリ107と物理ディスク装置105との間のデータ転送を行う。前記キャッシュメモリ107には、物理ディスク装置105の中のアクセス頻度の高いデータをロードしておく。このロード処理は、前記データレクタ106が実行する。ロードするデータの具体例は、データ処理装置100のCPU101のアクセス対象データや、このアクセス対象データと物理ディスク装置105上の格納位置に近いデータ等である。前記データレクタ108は、前記キャッシュメモリ107の管理情報を格納する。前記不揮発性メモリ109は、前記キャッシュメモリ107と同様に、物理ディスク装置105の中のアクセス頻度の高いデータをロードしておく。前記不揮発性メモリ管理情報110は、前記論理物理情報110の管理情報を格納する。前記論理物理情報110は、各論理ディスク装置105に配置されている論理ディスク装置（図2の200）を示す情報である。この情報は、データ処理装置100のCPU101のアクセス対象データと物理ディスク装置105上の格納領域の算出などを行う。前記論理物理情報400は、各論理ディスク装置（図2の200）のアクセス可否の状態を示す。前記アクセス情報500は、各論理ディスク装置（図2の200）のアクセス頻度やアクセスパターンなどの情報である。

【0023】論理物理情報300と論理物理情報400は、電源断などによる消失を防ぐために不揮発性の媒体に記録する。

【0024】前記物理ディスク装置105は、データを記録する媒体と、記録されたデータを読み書きする装置

とから構成される。

【0025】前記SVP111は、アクセス情報500の保守員への提示や、保守員からの再配置指示620の入力の受け付けを行う。また、保守員からの情報処理システム1への指示の発信や、情報処理システム1の障害状態等の保守員への提示を行う。

【0026】図2は、論理ディスク装置200と物理ディスク装置105の関連を表わした図である。論理ディスク装置200は、データ処理装置100のCPU101が演算アクセスする見掛け上のディスク装置で、アクセス対象データが実際に格納される物理ディスク装置105と対応している。論理ディスク装置200上のデータは、シーケンシャルアクセスを考慮して、物理ディスク装置105上に連続的に配置されている。論理ディスク装置200のデータが配置されている物理ディスク装置105がディスクアレイ構成の場合、論理ディスク装置200は複数の物理ディスク装置105に格納できる場合には、該物理ディスク装置105は複数の論理ディスク装置200と対応する。この論理ディスク装置200と物理ディスク装置105の対応は前記論理物理対応情報300で管理される。例えば、データ処理装置100のCPU101が論理ディスク装置200のデータ201をリードする時、記憶制御装置104で論理物理対応情報300に基づき論理ディスク装置200に対応する物理ディスク装置105を求め、その物理ディスク装置105の領域内のデータ格納位置202を求め、データ転送を行う。

【0027】図3は、論理物理対応情報300を表わした図である。論理物理対応情報300は、論理ディスク装置310と、物理ディスク装置320とから構成される。前記論理ディスク装置310は、各論理ディスク装置200が配置されている物理ディスク装置105上の領域に関する情報であり、論理ディスク装置200から対応する物理ディスク装置105を求める時に用いる。一方、前記物理ディスク装置320は、各物理ディスク装置105に配置されている論理ディスク装置200に関する情報で、物理ディスク装置105から対応する論理ディスク装置200を求める時に用いる。

【0028】前記論理ディスク装置310は、物理ディスク装置グループ311、RAID構成312および開始位置313の組を、論理ディスク装置200の数だけ持っている。前記物理ディスク装置グループ311は、当該論理ディスク装置200が配置されている物理ディスク装置105を示す情報である。前記RAID構成312は、前記物理ディスク装置グループ311のRAIDレベルを示す。前記開始位置313は、当該論理

ディスク装置200が物理ディスク装置105上で配置されている先頭位置を示す。

【0029】前記物理ディスク構成情報320は、論理ディスク装置グループ321を、物理ディスク装置105の数だけ有している。前記論理ディスク装置グループ321は、当該物理ディスク装置105に配置されている論理ディスク装置200を示す。

【0030】図4は、論理ディスク情報400を表わした図である。論理ディスク情報400は、論理ディスク状態401と再配置完了ポインタ402とを、論理ディスク装置200の数だけ有している。前記論理ディスク状態401は、「正常」「閉塞」「フォアマウント中」「再配置中」などの論理ディスク装置200の状態を表わす。前記再配置完了ポインタ402は、前記論理ディスク状態401が「再配置中」の時のみ有効な情報で、当該論理ディスク装置200の再配置処理を完了している領域の次の位置すなわち当該論理ディスク装置200が未だ再配置処理を終えていない領域の先頭位置を示す。「再配置中」におけるデータアクセス時、再配置完了ポインタ402よりも前の領域へのアクセスの場合には、再配置後の物理ディスク装置105へアクセスしなければならぬ。一方、再配置完了ポインタ402以後の領域へのアクセスの場合には、再配置前の物理ディスク装置105へアクセスしなければならない。

【0031】図5は、アクセス情報500を表わしている。アクセス情報500は、アクセス頻度情報501とアクセスパターン情報502とを、論理ディスク装置200の数だけ有している。このアクセス情報500は、記憶制御装置104、データ処理装置100、SVP111のいずれからも参照することが出来る。前記アクセス頻度情報501は、単位時間あたりの当該論理ディスク装置200へのアクセス回数を管理する。このアクセス頻度情報501は、各論理ディスク装置200の中でアクセス頻度の高いもの又は低いものを求める指標として用いる。前記アクセスパターン情報502は、当該論理ディスク装置200へのシーケンシャルアクセスとランダムアクセスの割合を管理する。このアクセスパターン情報502は、シーケンシャルアクセスが多く、よりシーケンシャル性能の高い物理ディスク装置105に再配置するの望ましい論理ディスク装置200を求める指標として用いる。

【0032】次に、記憶制御装置104の動作を説明する。図6は、記憶制御装置104の動作を詳細に表わした図である。まず、リード/ライト処理時の動作について説明する。ディレクタ106は、通常リード/ライト処理を実行する際、CPU101からチャネル103を経由してCPUからの指示600を受け取る。このCPUからの指示600は、リード（またはライト）対象のレコードが記憶されている論理ディスク装置200を指定する指定情報1と、リード（またはライト）対象のレ

コードが記憶されている論理ディスク装置200内の位置（トラック、セクタ、レコード）を指定する指定情報2とを含んでいる。ディレクタ106は、物理ディスク装置上のアクセス位置算出処理（610）で、前記CPUからの指示600と論理物理対応情報300とを用いて、物理ディスク装置105上でのアクセス位置を算出する。この物理ディスク装置アクセス位置算出処理（610）については図8を参照して後で詳述する。その後、たとえばリード処理では、算出した物理ディスク装置105上のデータ格納位置202のデータをキャッシュメモリ107上に読み上げてデータ201とし、その読み上げたデータ201をチャネル103を通じて主記憶102に転送する。

【0033】次に、アクセス情報500の採取処理について説明する。CPU101からのリード/ライト処理のアクセス時に、ディレクタ106は、アクセス対象論理ディスク装置200のアクセス情報500を更新する。アクセス頻度情報501の採取は、例えば、アクセスの度に内部カウンタをカウントアップしていき、一定時間または一定回数のアクセス経過後のアクセス時に、前記内部カウンタからアクセス頻度を判定する。アクセスパターン情報502の採取は、例えば、アクセスの度に内部カウンタにシーケンシャルアクセス回数をカウントアップしていき、一定時間または一定回数のアクセス経過後のアクセス時に、前記内部カウンタからアクセスパターンを判定する。

【0034】次に、再配置指示620を説明する。保守員は、SVP111を通じて提示されたアクセス情報500を参照して、各論理ディスク装置200の再配置の必要性を検討する。この検討の結果、再配置を決定した論理ディスク装置200があれば、SVP111を通じて記憶制御装置104に対して再配置指示620を出す。この再配置指示620は、再配置対象の論理ディスク装置200を2つ指定する指示情報1-2からなる。保守員が行う検討の内容は、後述する第3の実施形態で図10を参照して説明する論理ディスク装置再配置要求決定処理（910）と同様である。

【0035】次に、論理ディスク装置再配置処理（630）を説明する。ディレクタ106は、前記再配置指示620を受けて、指定された2つの論理ディスク装置200の間で論理ディスク装置再配置処理（630）を行う。図7は、論理ディスク装置再配置処理630の処理フロー図である。ステップ700では、論理ディスク情報400のうちの指定された2つの論理ディスク装置200の論理ディスク状態401を「再配置中」に設定更新する。ステップ701では、論理ディスク情報400のうちの指定された2つの論理ディスク装置200の再配置完了ポインタ402と各論理ディスク装置200の先頭位置に初期化する。ステップ702では、論理ディスク情報400のうちの指定された2つの論理ディスク装

置200の再配置完了ポインタ402をチェックし、全領域の再配置が完了していないればステップ703へ進み、完了していればステップ707へ進む。

【0036】ステップ703では、再配置完了ポインタ402が示すデータ位置から再配置処理10の処理単位分のデータに対して物理ディスク装置105からキャッシュメモリ107上へのデータ転送を行う。ここで、1回の処理単位分のデータ量は、再配置対象の2つの論理ディスク装置200の冗長データ1つに対応する各データ量の最小公倍数に決定される。たとえば、再配置対象のRAID5の論理ディスク装置200とRAID1の論理ディスク装置200の間で行うならば、RAID1の論理ディスク装置200の冗長データ1つに対応するデータ量は「1」であるから、1回の処理単位分のデータ量は、RAID5の論理ディスク装置200の冗長データ1つに対応するデータ量すなわちパリティ1つに対応するデータ量に決定される。

【0037】ステップ704では、再配置対象の各論理ディスク装置200の再配置先論理ディスク装置200がパリティを有するRAIDレベルのものである場合、キャッシュメモリ107上の再配置対象の1回の処理単位分のデータ201に対してパリティを生成する。ステップ705では、キャッシュメモリ107上の再配置対象の1回の処理単位分のデータ201および前記パリティの1回の処理単位分のデータ201を、再配置先の物理ディスク装置105へ書き込む。ステップ706では、1回の再配置処理単位分の再配置完了ポインタ402を進める。そして、前記ステップ702に戻る。

【0038】なお、上記ステップ703、704においては、データおよびパリティは、不揮発性メモリ109にも転送して二重化し、キャッシュ構造によるデータ消失を防ぐ。この理由は、上記ステップ705での書き込み時に、例えば、第1の論理ディスク装置200と第2の論理ディスク装置200のデータのうち、第1の論理ディスク装置200のデータを物理ディスク装置105（元は第2の論理ディスク装置200に配置されていた物理ディスク装置105）へ書き込んだ段階で障害によりキャッシュメモリ107上のデータがアクセス不能になったとすると、書き込みが終了していない第2の論理ディスク装置200のデータが消失するからである（元は第2の論理ディスク装置200に配置されていた物理ディスク装置105には、上記のように第1の論理ディスク装置200のデータが書き込まれてしまっている）。【0039】ステップ707では、論理物理対応情報300を更新する。すなわち、論理物理対応情報300と物理ディスク構成情報321とを変更する。ステップ708では、論理ディスク情報400の論理ディスク状態401を元の状態に戻し、再配置処理（630）を終了する。

【0040】次に、物理ディスク装置アクセス位置算出

処理(610)を説明する。図8は、物理ディスク装置アクセス位置算出処理部610の処理フロー図である。ステップ800では、論理ディスク情報400のうちのアクセス対象論理ディスク装置200の論理ディスク状態401が「再配置中」であるかをチェックし、「再配置中」ならばステップ801に進み、「再配置中」でなければステップ803に進む。

[0041] ステップ801では、論理ディスク情報400のうちのアクセス対象論理ディスク装置200の再配置完了ポイント402とアクセスデータ位置とを比較し、アクセスデータ位置が再配置完了ポイント402の指す位置以後ならばステップ802に進み、アクセスデータ位置が再配置完了ポイント402の指す位置より前ならばステップ803に進む。

[0042] ステップ802では、当該論理ディスク装置200の再配置完了の論理ディスク装置200をアクセス対象にする。そして、ステップ804へ進む。

[0043] ステップ803では、当該論理ディスク装置200をアクセス対象とする。

[0044] ステップ804では、アクセス対象の論理ディスク装置200に対応した物理ディスク装置105上でのアクセス位置を、論理物理対応情報300を用いて算出する。

[0045] 以上の第1の実施形態にかかる情報処理システム1および記憶制御装置104によれば、アクセス情報500に基づく保守員の判断により、アクセス頻度の高い論理ディスク装置をより高速な物理ディスク装置へ再配置することが出来る。また、シーケンシャルアクセスの比率の高い物理ディスク装置をよりシーケンシャルアクセスの比率の低い物理ディスク装置へ再配置することが出来る。従って、アクセス性能を向上することが出来る。

[0046] 一第2の実施形態一

上記第1の実施形態を变形して、記憶制御装置104からアクセス情報500をデータ処理装置100に提示し、データ処理装置100が再配置可否を決定し記憶制御装置104に再配置指示(620相当)を出すようにしてもよい。

[0047] 一第3の実施形態一

上記第3の実施形態は、再配置指示をSVPI11やデータ処理装置100から受けるのではなく、記憶制御装置104が自己決定するものである。

[0048] 図9は、記憶制御装置104の動作を詳細に表わした図である。第1の実施形態(図6)との違いは、論理ディスク再配置可否決定処理部910が再配置指示620を出すことである。

[0049] 図10は、上記論理ディスク再配置可否決定処理部910の処理フロー図である。この論理ディスク再配置可否決定処理(910)は、ディレクタ106が一周期で各論理ディスク装置200のアクセス情報

よい。信頼性を指標に用いれば、論理ディスク装置200上のデータの信頼性を向上させることが出来る。

[0057]

【発明の効果】本発明の記憶制御装置によれば、シーケンシャルアクセスの場合やランダムアクセスでヒット率が低い場合でも、アクセス性能を向上することが出来る。また、本発明の記憶制御装置によれば、データの信頼性を向上することが出来る。

【図面の簡単な説明】

図1 本発明の第1の実施形態にかかる記憶制御装置を含む情報処理システムのプロック図である。

図2 論理ディスク装置と物理ディスク装置との対応関係の説明図である。

図3 論理物理対応情報の構成例示図である。

図4 論理ディスク情報の構成例示図である。

図5 アクセス情報の構成例示図である。

図6 本発明の第1の実施形態における記憶制御装置の動作を示すプロック図である。

図7 論理ディスク装置再配置処理部の処理フロー図である。

図8 物理ディスク装置アクセス位置算出処理部の処理フロー図である。

図9 本発明の第3の実施形態における記憶制御装置の動作を示すプロック図である。

図10 論理ディスク装置再配置可否決定処理部の処理フロー図である。

【図3】

論理物理対応情報  
300

図3

論理ディスク構成情報 310



物理ディスク構成情報 320













- ・ 動的な領域管理の回避
- ・ ソートのチューニング
- ・ チェックポイント・アクティビティのチューニング
- ・ LGWR および DBWR の I/O のチューニング
- ・ バックアップおよびリストア操作のチューニング
- ・ 大規模プールの構成

## I/O の分散によるディスク競合の低減

この項では、ディスク競合を低減する方法を説明します。

- ・ ディスク競合とは
- ・ データ・ファイルと REDO ログ・ファイルの分離
- ・ 表データのストライプ化
- ・ 表と索引の分離
- ・ Oracle と関係のないディスク I/O の低減

### ディスク競合とは

複数のプロセスが同時に同じディスクにアクセスしようとするとき、ディスク競合が発生します。多くのディスクには、アクセス数と 1 秒あたりに転送できるデータ量の両方について制限があります。これらの制限に達すると、ディスクにアクセスするためにプロセスを待機させることが必要になります。

通常は、V\$FILESTAT ビューの統計およびオペレーティング・システムの機能を検討してください。ディスク性能の限界を把握するために、ハードウェアのマニュアルを調べてください。最大性能やそれに近い性能で動作しているディスクはディスク競合の候補です。たとえば、VMS や UNIX オペレーティング・システム上の一部のディスクでは、1 秒あたり 60 以上の I/O は過剰となる場合があります。

また、db file sequential read、db file scattered read、db file single write、db file parallel write のイベントについて、V\$SESSION\_EVENT を見なおしてください。これらはすべて、データ・ファイル・ヘッダー、コントロール・ファイルに対して実行される I/O に対応したイベントです。これらの待機イベントのいずれかが高い平均時間を示している場合は、sar または iostat を使用して I/O 競合を調査してください。次に、デバイスでのビジー待機を探します。ファイル統計を検討して、どのファイルが高い I/O と関連しているかを特定します。

負荷が過剰になっているディスクに対するアクティビティを削減するには、そのディスク上にあるアクセス頻度の低い 1 つ以上のファイルを、それほど負荷のないディスクに移動します。すべてのディスクの I/O 量がだいたい同じになるまで、各ディスクにこの原則を適用してください。これは、I/O 分散と呼ばれます。

## データ・ファイルと REDO ログ・ファイルの分離

Oracle プロセスは、絶えずデータ・ファイルと REDO ログ・ファイルにアクセスします。これらのファイルが同じディスク上に存在している場合、ディスク競合が発生する可能性があります。各データ・ファイルを別々のディスク上に配置してください。そうすると、複数のプロセスがディスク競合せずに同時に異なるファイルにアクセスできます。

REDO ログ・ファイルの各セグメントは、他のアクティビティがない別々のディスクに配置してください。REDO ログ・ファイルは、トランザクションがコミットされると、ログ・ライター・プロセス (LGWR) によって書き込まれます。REDO ログ・ファイル内のセグメントは順次書き込まれます。同じディスクに対する同時実行のアクティビティが存在しない場合、この順次書き込みはさらに高速で行われる可能性があります。REDO ログ・ファイルに別々の専用のディスクを割り当てると、さらにチューニングしなくても通常は LGWR が円滑に実行されます。LGWR に関連するパフォーマンス上のボトルネックはめったにありません。

関連項目： LGWR のチューニングの詳細は、21-14 ページの「REDO ログ・バッファ・ラッチの競合の検出」を参照してください。

データ・ファイル専用ディスクを用意することと REDO ログ・ファイルをミラー化することとは、重要な安全対策です。これらの手順を実行することによって、データ・ファイルと REDO ログ・ファイルの両方を単一のディスク障害で失う可能性がないことが保証されます。REDO ログ・ファイルのミラー化によって、REDO ログ・ファイルを単一のディスク障害で失う可能性はないことが保証されます。

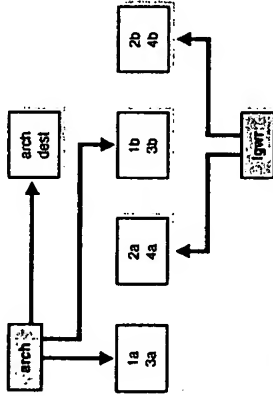
アーカイバ・プロセスと LGWR (マルチ・メンバー・グループを使用している場合) 間の I/O 競合を防ぐには、アーカイバの読み込みと LGWR の書き込みが別個に実行されることを確認してください。たとえば、システムに 2 つのメンバーを持つグループが 4 つある場合、次のシナリオを使用してディスク・アクセスを分離してください。

4 グループ x 2 メンバー = 8 ログ・ファイルで、ラベルは 1a、1b、2a、2b、3a、3b、4a、4b。それには、最低でも 4 つのディスクと、アーカイバ・ファイル用に 1 つのディスクが必要で

す。

図 20-2 は、競合を最小限にするために、ディスク間で REDO メンバーを分散する方法を示しています。

図 20-2 ディスク間で REDO メンバーの分散



この例では、LGWR はログ・グループ 1 (メンバー 1a と 1b) を切替えて外し、現在はログ・グループ 2 (2a と 2b) に書込みを行っています。同時に、アーカイブ・プロセスは、グループ 1 から読込みをして、アーカイブ先へ書込みを行っています。REDO ログ・ファイルがどのような順序で分散されているかに注意してください。

注意: REDO ログ・ファイルをミラー化する、すなわち各 REDO ログ・ファイルの複数のコピーを別々のディスク上に保持することで、LGWR が大幅に遅くはなりません。LGWR は、各ディスクに対して並列して書込みを行い、並列書込みの各部分が完了するまで待機します。単一のディスク書込みを実行するために必要な時間は変動することがあるため、コピーの数が増えると、並列書込みでの単一のディスク書込みにかかる時間が平均よりも低くなる傾向が顕著です。ただし、並列書込みが、最も長い単一のディスク書込みよりも長くなることはありません。また、並列した書込みに関連するオーバーヘッドがオペレーティング・システムで多少発生する場合があります。

## 表データのストライプ化

ストライプ化、すなわち大きな表のデータを別々のディスク上の別々のデータ・ファイルに分散させることも、場合の低減に役立ちます。

関連項目: この方針については、20-21 ページの「ディスクのストライプ化」で詳しく説明します。

## 表と索引の分離

頻繁に使用される表は、索引と分離する必要があります。一連のトランザクション中は、索引が最初に読み込まれてから表が読み込まれます。これらの I/O は順次に発生するので、

表と索引を同じディスク上に格納しても問題は発生しません。ただし、非常に高度な OLTP システムでは、索引と表の分離が必要な場合があります。

索引と表を個別の表領域に分離して、ディスク・ヘッドの移動を最小限に留め、I/O をパラレル化します。すると、1 つのディスク・ヘッドが索引データ上に、残りのヘッドが表データにあるので、両読込みともに高速になります。

同時にアクセスされるオブジェクトを分離するという考え方は索引にも当てはまります。たとえば、SQL 文が 2 つの索引を同時に使用する場合、別々のディスクに索引があればパフォーマンスは改善されます。

また、同じディスクに頻繁にアクセスされる複数の表を配置することは避けてください。これを行うには、アプリケーション・アクセス・パターンを熟知している必要があります。

パーティション表とパーティション索引を使用すると、データ・ウェアハウスの操作のパフォーマンスを改善できます。大きな表あるいは索引を異なる表領域に配置する複数の物理セグメントに分割します。大きなオブジェクト・データ・タイプのある表もすべて個別の表領域に配置してください。

## Oracle と関係のないディスク I/O の低減

可能であれば、データベース・ファイルを含むディスクについて、Oracle と関連のない I/O を取り除いてください。この処理は、REDO ログ・ファイルへのアクセスを最適化することで有効です。これによってディスク観測が減少するだけでなく、動的パフォーマンス表 vs FILESTAT を使用して、そのようなディスク上のアクティビティをすべて監視することもできます。

## ディスクのストライプ化

この項では、次のトピックについて説明します。

- ストライプ化の目的
- I/O のバランス化とストライプ化
- ディスクを手動でストライプ化する方法
- オペレーティング・システム・ソフトウェアでディスクをストライプ化する方法
- ストライプ化と RAID

## ストライプ化の目的

ストライプ化によって、大きな表のデータが小さな部分に分割され、これらの部分が別々のディスク上の別々のデータ・ファイルに格納されます。これによって、複数のプロセスがディスク観測なしで表の異なる部分に同時にアクセスできます。ストライプ化は、数多くの行を持つ表へのランダム・アクセスを最適化する上で特に有効です。ストライプ化は、手動で実行する (以降で説明します) ことも、オペレーティング・システムのストライプ化ユーティリティを使用して実行することもできます。

## IO のバランシングとストライプ化

これまで、ベンチマークのチューニング担当者は、使用可能なデバイス上で I/O の負荷のバランシングを均一にすることを懸命に試行してきました。根拠は、オペレーティング・システムによって、頻繁に使用されるコンテナ・ファイルを多数の物理デバイスにストライプ化する機能が提供されています。ただし、このような手法は、負荷の再分散によってなんらかの形態のキューが排除または削減される場合のみ有効です。

ドライブでの高ビジー率とともに待機サービス時間が存在する場合には、I/O の分散が必要となります。多数の物理ドライブを使用可能な場合は、2 つの専用ドライブで REDO ログをとることを検討してください。一つである理由は、REDO ログはオペレーティング・システムまたは Oracle REDO ログ・グループ機能を使用して常にミラー化する必要があるからです。REDO ログはシリアルに書き込まれるので、REDO ログ・アクティビティ専用のドライブでは過剰なヘッドの移動は避けられます。このため、ログ書き込みのスピードが大幅に向上します。

アーカイブする場合は、LGWR および ARCH が同じ書き込み / 読み込みヘッドを競合しないように、別のディスクを使用することが効果的です。これは、ログを代替ドライブに配置することで行います。

ミラー化は、I/O ボトルネックの原因となる可能性もあります。各ミラーへの書き込みプロセスは、通常は並列で行われるので、ボトルネックの原因にはなりません。ただし、各ミラーが別々にストライプ化されている場合は、低速のミラー・メンバーが壊れるまで I/O は完了しません。I/O の問題を回避するためには、対象データベース（つまりコピー）で元データベースと同等のディスクを使用してストライプ化を行ってください。

たとえば、8 頭のディスクに 160KB のデータがストライプ化し、データが 1 つのディスクにしかミラー化されていない場合は、データが 8 頭のディスク上でどれだけ高速に処理されるかにかかわらず、160KB がミラー・ディスクに書き込まれるまで I/O は完了しません。したがって、データベースへの書き込みには 20.48 ミリ秒しかかかりませんが、ミラーへの書き込みには 137 ミリ秒を要します。

## ディスクを手動でストライプ化する方法

ディスクを手動でストライプ化するには、オブジェクトの記憶域要件を I/O 要件と関連付ける必要があります。

1. 最初に、次の項目を調べて、オブジェクトのディスク記憶域要件を評価します。
  - ・ オブジェクトのサイズ
  - ・ ディスクのサイズ

たとえば、オブジェクトが 5GB の Oracle 記憶域を必要とする場合は、それを収容する 1 つの 5GB のディスクまたは 2 つの 4GB のディスクが必要です。一方、システムが 1GB または 5GB のディスクで構成されている場合は、オブジェクトはそれぞれ 5 個または 3 個のディスクを必要とすることがあります。

2. 20-3 ページの「I/O 要件の分析」で説明したアプリケーションの I/O 要件とこれと比較してください。記憶域要件と I/O 要件の大きい方をとる必要があります。

たとえば、記憶域要件が 5 つのディスク（それぞれ 1GB）であり、I/O 要件が 2 つのディスクである場合は、アプリケーションは大きい方の側である 5 ディスクを必要とします。

3. CREATE TABLESPACE 文で記憶域を作成します。DATAFILE 句にデータ・ファイルを指定します。各ファイルは異なるディスク上に作成してください。次に例を示します。

```
CREATE TABLESPACE stripedbspace
DATAFILE 'file_on_disk_1' SIZE 1GB,
        'file_on_disk_2' SIZE 1GB,
        'file_on_disk_3' SIZE 1GB,
        'file_on_disk_4' SIZE 1GB,
        'file_on_disk_5' SIZE 1GB;
```

4. CREATE TABLE 文で表を作成します。TABLESPACE 句に新たに作成した記憶域を指定します。

また、STORAGE 句に表のエクステンツのサイズを指定します。別々のデータ・ファイルに各エクステンツを格納します。表のエクステンツは、オーバーヘッドを考慮して表領域内のデータ・ファイルより小さくしてください。たとえば、1GB (1024MB) のデータ・ファイルを準備するときは、表エクステンツを 1023MB に設定できます。次に例を示します。

```
CREATE TABLE stripedtab (
  col_1 NUMBER(2),
  col_2 VARCHAR2(10) )
TABLESPACE stripedbspace
STORAGE ( INITIAL 1023MB NEXT 1023MB
MINEXTENTS 5 PCTINCREASE 0 );
```

(あるいは、DATAFILE 'd:\offld' SIZE 'size' 句を指定した ALTER TABLE ALLOCATE EXTENT 文を入力することで表をストライプ化することもできます。)

これらのステップによって、表 STRIPEDTAB が作成されます。STRIPEDTAB には、それぞれサイズが 1023MB の初期エクステンツが 5 つあります。各エクステンツは、CREATE TABLESPACE 文の DATAFILE 句に指定されたデータ・ファイルの 1 つを取り上げます。これらのファイルはすべて別々のディスク上に存在します。MINEXTENTS が 5 なので、これら 5 つのエクステンツはすべて同時に割り当てられます。

関連項目: MINEXTENTS および他の記憶域パラメータの詳細は、「Oracle8i SQL リファレンス」を参照してください。

## オペレーティング・システム・ソフトウェアでディスクをストライプ化する方法

手動でディスクをストライプ化する方法のかわりとして、LVM (論理ボリューム・マネージャ) などのオペレーティング・システム・ユーティリティやサード・パーティ・ツールを

使用したり、ハードウェア・ベースのストライプ化機能を使用して、ディスクをストライプ化します。

ユーティリティあるいはハードウェア・ベースのストライプ機構を使用する場合に考慮する主要要因は、ストライプ・サイズ、(ストライプ幅を定義する) ストライプ対象のディスク数および同時実行性のレベル (あるいはI/O アクティビティのレベル) です。これらの要因は、Oracle ブロック・サイズとデータベース・アクセス方法の影響を受けます。

表 20-14 最小ストライプ・サイズ

ディスク・アクセス	最小ストライプ・サイズ
ランダム読みおよび書き込み	最小ストライプ・サイズは、Oracle ブロック・サイズの2倍です。
順次読み込み	最小ストライプ・サイズは、DB_FILE_MULTIBLOCK_READ_COUNT の値の2倍です。

表 20-15 一般ストライプ・サイズ

同時実行性	IO サイズ	一般ストライプ・サイズ
低	小	$k \cdot \text{DB\_BLOCK\_SIZE}$
低	大	$k \cdot \text{DB\_BLOCK\_SIZE}$
高	小	$k \cdot \text{DB\_BLOCK\_SIZE}$
高	大	$k \cdot \text{DB\_BLOCK\_SIZE} \cdot \text{DB\_FILE\_MULTI\_BLOCK\_READ\_COUNT}$

ここで、 $k=2,3,4...$  です。

ストライプ化では、データへの統一的なアクセスが前提とされています。ストライプ・サイズが大きすぎる場合は、1つまたは少数のディスクでホット・スポットが発生する可能性があります。これは、ストライプ・サイズを小さくし、データをより多くのディスクに分散することで問題できます。

固定サイズの 100 行が 5 つのディスクに均等に分散され、各ディスクが 20 の順次行を含んでいる例を考えます。アプリケーションが行 35 ～ 55 へのアクセスのみを必要とする場合は、2 つのディスクのみですべての I/O をする必要があるため、同時実行性が高い場合には、システムは目的のパフォーマンス・レベルを達成できない場合があります。

この問題は、行 35 ～ 55 をより多くのディスクに分散することで解決できます。実行の例では、1 ブロックあたりに 2 行とすれば、行 35 と 36 が同じディスク上に存在し、行 37 と行 38 は別のディスクに存在することになります。このアプローチをとると、データはすべてのディスクに分散され、I/O スループットが改善されます。

## ストライプ化と RAID

RAID (Redundant arrays of inexpensive disks) 構成では、データの信頼性が改善されます。ただし、I/O パフォーマンスは実装されている RAID 構成によって異なります。

以下に、最も広く使用されている RAID 構成を示します。

- RAID 1: 信頼性および読み込み率が向上します。ただし、書き込みは不経済になります。
- RAID 0+1: 信頼性が向上し、RAID 1 よりも読み込みと書き込みのパフォーマンスが向上します。
- RAID 5: 高度の信頼性を提供します。順次読み込みを行うと最も多くの利点が引き出せます。書き込みパフォーマンスは RAID 5 ではかなり悪くなります。この構成は、書き込み量の多いアプリケーションには推奨できません。

注意: RAID 0 は読み込みおよび書き込みともに最高のパフォーマンスを提供しますが、冗長性がないため、突如には RAID システムではありません。Oracle では、RAID 0 システムに実用データベース・ファイルを配置しないようお勧めします。

最適なストライプ・サイズは次の 3 項の関数になります。

1. 配列に対する I/O 要求のサイズ
2. 配列に対する I/O 要求の同時実行性
3. ブロック・サイズ境界と一致する物理的なストライプ境界

ストライピングは、配列内の 2 つ以上のディスクへの I/O アクセスを分散させるには優れたツールです。ただし、次のテクニックを覚えておいてください。

- 同時実行性が高い配列では、単一 I/O 要求が複数の物理 I/O コールに分解されないことを確認する必要があります。そうでなければ、システムで実行される物理 I/O 要求数が向いにもなり、システム I/O 応答時間が大幅に下がります。
- 同時実行性の低い配列では、単一 I/O が同じディスクに 2 度アクセスすることがないようにする必要があります。同じディスクに 2 度アクセスすると、前述と同じパフォーマンス面でペナルティが発生します。

## 動的な領域管理の回避

表やローカルバック・セグメントのようなオブジェクトを作成すると、データのために領域がデータベース内に割り当てられます。この領域をセグメントと呼びます。後のデータベース操作によってデータ容量が増大し、割り当てられた領域を上回るようになると、Oracle はそのセグメントを拡張します。この場合、動的拡張によってパフォーマンスが低下します。

この項では、次のことについて説明します。